

# Conversational Financial Information Retrieval Model (ConFIRM)

Stephen Choi\*, William Gazeley\*, Siu Ho Wong, Tingting Li

LORA Research, 2A, 128 Wellington Street, Central, Hong Kong  
{stepchoi, william, adrian, tingting}@asklora.ai

## Abstract

With the exponential growth in large language models (LLMs), leveraging their emergent properties for specialized domains like finance merits exploration. However, regulated fields such as finance pose unique constraints, requiring domain-optimized frameworks. We present ConFIRM, an LLM-based conversational financial information retrieval model tailored for query intent classification and knowledge base labeling. ConFIRM comprises two modules: 1) a method to synthesize finance domain-specific question-answer pairs, and 2) evaluation of parameter efficient fine-tuning approaches for the query classification task. We generate a dataset of over 4000 samples, assessing accuracy on a separate test set. ConFIRM achieved over 90% accuracy, essential for regulatory compliance. ConFIRM provides a data-efficient solution to extract precise query intent for financial dialog systems.

## 1 Introduction

Advancements in large language models have seen extraordinary growth (Zhao et al, 2023) since language models demonstrated power scaling laws on cross-entropy loss (Kaplan et al, 2020). Similar efforts have also been made to leverage LLMs to improve information retrieval (IR) capabilities (Zhu et al, 2023). The next step in this evolution should focus on developing optimized architectures and training techniques to adapt these powerful models for specialized domains.

However, mission critical domains such as healthcare, legal, or *finance* present challenges beyond general AI optimization and alignment. Financial services are mission critical because they are regulated. For example, in financial services, approval from regulators like the Financial Industry Regulatory Authority (FINRA) is required for all data feeds.<sup>1</sup> For financial service providers, the ideal solution would comply with the regulatory requirements while also capturing the emergent properties of LLMs (Wei et al, 2022) like conversational knowledge base question answering (KBQA) for non-professional retail investors. In this paper, we examine ConFIRM (Conversational Financial IR Model), an LLM-based framework specifically tailored for high precision conversational financial information retrieval from a specialized financial Knowledge Base (KB).

ConFIRM is designed to precisely map natural language queries to the appropriate KB entries, while also maintaining high accuracy for queries that fall outside of the internal KB - essential for compliance with data distribution regulations. We focus on both internal/external KB labels as well as data category and field classifications; the full IR instructions fall outside the scope of this paper. We do not include the full IR instructions because data category and field classifications sufficiently meet the regulatory and majority of the business requirements. We focus on classification rather than the full retrieval instructions to highlight the practical business case, e.g., the “retrieval” components of Retrieval Augmented Generation (RAG)<sup>2</sup>, rather than additional engineering to handle units, currency, and other numeraire issues.

---

<sup>1</sup><http://www.finra.org/filing-reporting/trace/content-licensing/retransmission-data-feed-policy>

\*Equal contribution

<sup>2</sup><https://ai.meta.com/blog/retrieval-augmented-generation-streamlining-the-creation-of-intelligent-natural-language-processing-models>

Our approach diverges from the broader use case from few shot or no shot learning (Yu et al, 2021; Gao et al, 2023) as we observed that training sample sizes below 100 yielded substandard accuracy. Furthermore, we aim to offer more automated and economical solution than large-scale KQBA (Bordes et al, 2015). The first practical challenge lies in constructing a test and training set *ex nihilo*. Our solution is synthesize question-answer pairs utilizing Large Language Models (LLMs). These synthetically generated pairs serve as a finance domain-specific KBQA dataset for both training and evaluation purposes.

Given the specificity and structured nature of the KB space and heightened regulatory thresholds, the query intent (QI) classification (Broder 2002) for IR holds particular significance. For this task, we adopt parameter-efficient-fine-tuning (PEFT) methodologies (Lialin et al, 2023) on an LLM. This approach addresses a key requirement in practice: having direct access and control of the LLM to manage the proprietary nature of financial services data.

In this paper, we outline the two modules of ConFIRM: 1) a method to synthetically generate a finance domain-specific KBQA dataset of sufficient size and 2) evaluation on the data efficiency of PEFT methodologies for regulatory compliance. ConFIRM, designed for practical application in the finance domain, operates with minimal human intervention, requiring only limited supervision to validate external KB labels and manually generating seed tasks. To aid reproduction, we post our model code at: <https://github.com/WilliamGazeley/ConFIRM>

## 2 Methodology

In this section, we provide a formal definition of the conversational financial information retrieval task and detail the architecture of ConFIRM designed to solve this problem.

### 2.1 Knowledge Base

The sheer volume of financial data can be overwhelming, often amassed in terabytes. However, financial data within regulatory-approved knowledge bases must adhere to

structured or semi-structured formats: the Financial Data Transparency Act of 2022 (FDTA) requires all information reported to financial regulators to be electronically searchable<sup>3</sup>.

Our internal knowledge base structure is modeled after **Refinitiv Datastream**, a leading and widely used provider of financial data. Datastream offers a comprehensive range of global financial data from various sources, including stock exchanges, central banks, and other financial institutions. The platform also provides information on the data category, data types, and descriptions of each data field. Using this foundation, we structure the following data categories for a hypothetical stock investment firm. Our focus is primarily on stocks since stock is the most prevalent form of active retail investment; as reported by Gallup, 61% of Americans are stock owners<sup>4</sup>. This hypothetical framework can be easily extended for additional or different data categories. (For detailed information on the actual fields, please refer to Appendix A.)

**Stock data:** This category is a table that contains data about individual stocks, including price data, identifiers, descriptions, and data available in company filings. For the scope of this paper, we limit the number of data fields to the top 40 ranked by popularity within the category marked as “Equities” by Refinitiv.

**Market data:** This category holds data about the overall stock market, including benchmark indices such as S&P 500, and sector and style benchmarks. We limit the number of data fields to the top 15 ranked by popularity within the “Equity Indices” category.

**Economic data:** This category holds data and statistics about the US economy as a whole, such as GDP, unemployment rate and consumer price index as well as leading indicators. We limit the number of data fields to the top 11 ranked by popularity within the “Economics” category.

**News data:** This category is a data warehouse where all news scraped from Reuters (one of the largest news agencies) is stored.

**External data:** This category represents the external knowledge base where all data that is out-of-scope of the four internal categories. In terms of internal/external KB label accuracy, this category

<sup>3</sup> James M. Inhofe National Defense Authorization Act for Fiscal Year 2023, P.L. 117-263 (2022)

<sup>4</sup><https://news.gallup.com/poll/506303/stock-ownership-highest-2008.aspx>

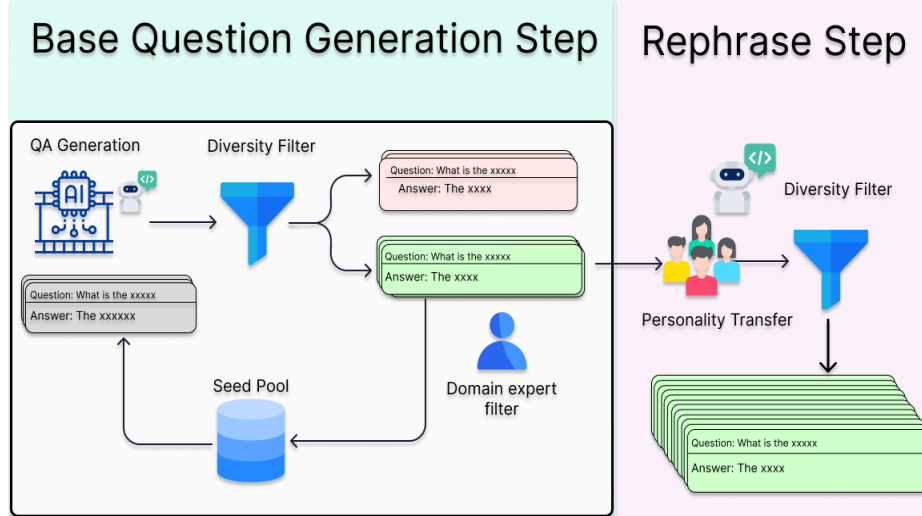


Figure 1: A schematic illustration of synthetic question generation

represents the “true negatives”. A typical example of this data source is social media or internet sites.

## 2.2 QA generation

To generate natural sounding questions that align with ConFIRM’s financial KB, we adopt the iterative instruction generation approach of SELF-INSTRUCT (Wang et al, 2023). It is critical to synthesize colloquial, genuine user interactions. For this we leverage an Extra Large Language Model (XLLM) such as PaLM 2 (Google, 2023) and Chat GPT 3.5 (Brown et al, 2020), each with over 100 billion parameters to generate “conversational” questions.

We apply an LLM (e.g., PaLM2 and GPT-3.5) to generate new questions (zero-shot) with a prompt “generate questions that can only be answered by the following data label:” serving as the seed task instruction, where the data category, data field name, description, and potential multi-field permutations are provided as the output (answers). As all tasks involve classification, we use the Output-first Approach of the SELF-INSTRUCT framework. This involves initially assigning the class labels (category and field) and then conditioning the input (question) generation based on each pre-defined label. A domain expert selects the best questions and re-seeds by appending the sample questions to the seed pool.

Through an iterative process of identifying the best questions for seeding (conducted four times in our application), manually filtering for invalid

samples, we obtain 1000 high quality question-answer pairs. To further emulate a conversational tone, we utilize the Text2Text generation method proposed by Ramirez et al, 2023. We employ a few-shot prompt built with a dataset based on the big five (OCEAN) personality traits (McCrae and Costa, 1999) to incorporate diverse personality styles onto the original 1000 questions. Pseudo-references, derived from meaning representations, are passed to an LLM (e.g., PaLM2, GPT 3.5) for few-shot in-context learning to generate questions in an extraverted personality style.

The final set undergoes filtering for ROUGE-L (Lin and Och, 2004) scores both vertically (across the questions) and horizontally (within individual question-answer pairs), to eliminate repetition, reduce trivial assignments, and promote diversity. This comprehensive process (Figure 1) results in over 3300 samples with a mere 10 hours of human intervention. We allocate 3000 samples for training, 300 for validation (10% additional to the training). To reduce data snooping, we generate a separate additional 1000 samples following the same process above for testing samples. Refer to Appendix B for examples.

## 2.3 Parameter Efficient Fine-Tuning (PEFT)

With the training and test sets, we focus on parameter efficient fine tuning on a manageable pre-trained LLM, Llama-2-7b. This arrangement allows full control of the model location, network,

Hyperparameter	LoRA	P-Tuning	Adapter
learning rate	3E-04	3E-03	1E-02
batch size	8	3	4
epochs	50	50	50
max length	128	128	128
r	4		
dropout	1.00E-03		
alpha	64		
encoder hidden size		128	
number of virtual tokens		15	
length			10
layers			8

Table 1: PEFT method hyperparameters

and hardware – factors that are essential for certain financial service providers. From a technical perspective, preserving the weights of the pre-trained Llama-2-7b ensures that the footprint is sufficiently compact for practical applications in environments with limited resources.

The goal is to convert a general-purpose base model (Llama-2-7b) to a specialized ConFIRM classifier. We retain the weights of the pre-trained models and apply established PEFT methods; the code for our PEFT is based on the PEFT library provided by HuggingFace<sup>5</sup>. Three established methods were evaluated for data efficiency: adapters (Zhang et al, 2023), p-tuning (Liu et al, 2023), and Low Rank Adaptation (Hu et al, 2021). This set encapsulates pre-transformer layer additive PEFT method, post-transformer layer additive PEFT method, and a reparameterization PEFT method. For each of these PEFT methods, we utilize the validation set to optimize the hyperparameters (Table 1). These include the standard hyperparameters such as learning rate and batch size, as well as method-specific hyperparameters such as Low Rank Adaptation (LoRA) scaling factor, the number of adapter layers, and the number of tokens.

### 3 Results

The accuracy of the ConFIRM model is evaluated on two factors: accuracy of internal/external KB labeling and precision of data category and field

labels. Instances where false positives or false negatives occurred in external labels are marked as failed classification, underscoring regulatory concerns. A classification is determined as successful if the ConFIRM model returned at least a superset of the correct data category or categories within the answer set. For each PEFT method, the optimal hyperparameters are chosen from validation testing and then the optimal configuration is used to test for data efficiency.

The most notable finding is the markedly superior performance of LoRA over other additive PEFT methods (Figure 2). LoRA, a reparameterization-based PEFT method, was able to reach 88% accuracy with 1500 training samples and surpassed 91% with 3000 training samples. In contrast, both additive-based PEFT methods failed to exceed 65%, and adapter methods were limited to 33% even with 3000 training samples. The results (Table 2) reveal two clear outcomes among the PEFT methods: 1) reparameterization of all layers, even with a relatively low rank (4), yielding 2,097,152 trainable parameters – a mere 0.031% of the original LLM - produces the best performance and 2) p-tuning layers added before the transformer layers are more effective than adaptive layers added after the transformer layers. In line with the findings of Lester et al. (2021), which demonstrated that prompt tuning became more

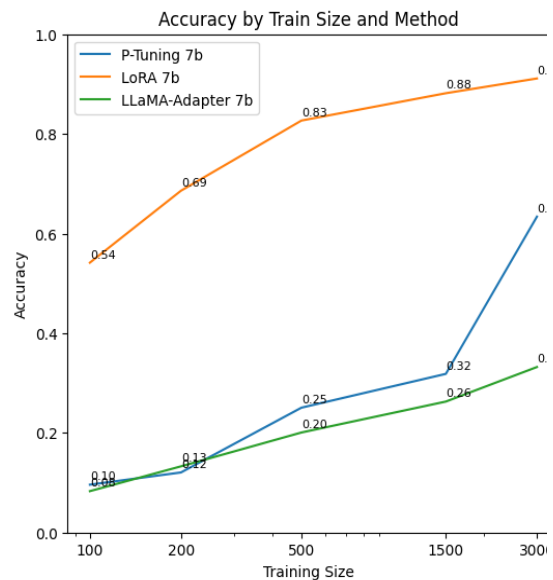


Figure 2: Data efficiency of the PEFT methods

<sup>5</sup><https://github.com/huggingface/peft>

Method	Training Set Size	Inference Runtime (seconds)	Trainable Parameters	Accuracy
P-Tuning	100	0.74	1,130,752	9.6%
P-Tuning	200	0.94	1,130,752	12.1%
P-Tuning	500	0.85	1,130,752	25.1%
P-Tuning	1500	0.77	1,130,752	31.8%
P-Tuning	3000	0.75	1,130,752	63.4%
P-Tuning*	3000	1.36	1,409,280	70.0%
LoRA	100	0.59	2,097,152	54.2%
LoRA	200	0.58	2,097,152	68.6%
LoRA	500	0.60	2,097,152	82.7%
LoRA	1500	0.60	2,097,152	88.2%
LoRA	3000	0.61	2,097,152	91.1%
Adapter	100	1.26	327,688	8.3%
Adapter	200	1.17	327,688	13.3%
Adapter	500	0.75	327,688	20.1%
Adapter	1500	1.25	327,688	26.3%
Adapter	3000	1.93	327,688	33.2%

Table 2: PEFT method results  
\* Llama-2-13b

effective with model parameters sizes above 10 billion, we experimented with larger LLMs for prompt tuning. However, prompt tuning to the Llama-2-13b model with 13 billion parameters using 3,000 training samples, still failed to exceed 70% accuracy.

The subpar performance of p-tuning and adapter PEFT methods is especially disappointing from the perspective of runtime efficiency. The runtime for classification (inference) on each test questions for LoRA-tuned model on a single Nvidia A100 GPU is 0.61 seconds while p-tuned model runs for 0.75 seconds and the adapter implementation requires 1.93 seconds. Longer runtimes can be attributed to the additional computation required by the extra p-tuning layers or adapter layers, further highlighting the advantages of a LoRA tuned model.

## 4 Conclusions

In this paper study we present ConFIRM - a conversational financial information retrieval framework that leverages the latest advances in large language models. We employ iterative prompt engineering and controlled paraphrasing to curate a domain-specific dataset for finance KBQA, *ex nihilo*. Various PEFT methods were examined, with the LoRA PEFT method proving particularly effective in enhancing the accuracy of

stock-investing KBQA frameworks beyond 90%, conforming to regulatory standards. While further research is necessary to optimize other PEFT methods and tackle the limitations present in our current study, the ConFIRM model offers a pragmatic approach to precise query intent classification within financial dialog systems. This is integral for both regulatory compliance and meeting business requirements. Moreover, our findings emphasize the promising potential of utilizing large language models within the financial sector, thus paving the way for novel applications of artificial intelligence in this domain.

## Limitations

While ConFIRM has achieved over 90% accuracy with LoRA implementation, we identify two major limitations to the current iteration of ConFIRM. The first major limitation is the incompleteness of the IR process. Though we have successfully created a comprehensive question-answer sample set, we have not subjected the system to comprehensive information retrieval testing. We anticipate that achieving this would require not only additional LLM modules to generate the precise instructions but also measurement enhancements since success or failure may not be as binary as query intent. It would likely necessitate more intricate engineering to address numeraire

issues and syntax. We plan to address this issue in future iterations of the model.

Second, we have not yet explored the application of this work beyond a 7 (and 13) billion parameter model. Both smaller and larger LLMs can have different setup requirements and data efficiencies, aspects which we plan investigate in future work.

## Acknowledgments

The authors would like to thank King Yeung Shum, Young Jin Kim, Nicolas Cheung, and other colleagues at LORA Research for valuable feedback. The computational work for this article was partially performed on resources at Google Cloud Platform with support by Google.

## References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin

Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, Yonghui Wu2023. [PaLM 2 Technical Report](#). *arXiv preprint arXiv:2305.10403*.

Antoine Bordes, Nicolas Usunier, Sumit Chopra and Jason Weston. 2015. [Large-scale Simple Question Answering with Memory Networks](#). *arXiv preprint arXiv:1506.02075*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever and Dario Amodei. 2020. [Language Models are Few-Shot Learners](#). *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Luyu Gao, Xueguang Ma, Jimmy Lin and Jamie Callan. 2023. [Precise Zero-Shot Dense Retrieval without Relevance Labels](#). *ACL 2023*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang and Weizhu Chen. 2021. [LoRA: Low-Rank Adaptation of Large Language Models](#). *arXiv preprint arXiv:2106.09685*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu and Dario Amodei. 2020. [Scaling Laws for Neural Language Models](#). *arXiv preprint arXiv:2001.08361*.

Brian Lester, Rami Al-Rfou and Noah Constant. 2021. [The Power of Scale for Parameter-Efficient Prompt Tuning](#). *EMNLP 2021*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel and Douwe Kiela. 2021. [Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](#). *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Xiang Lisa Li and Percy Liang. 2021. [Prefix-Tuning: Optimizing Continuous Prompts for Generation](#). *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4582–4597, Online. Association for Computational Linguistics.

Vladislav Lialin, Vijeta Deshpande and Anna Rumshisky. 2023. [Scaling Down to Scale Up: A](#)

- Guide to Parameter-Efficient Fine-Tuning. *arXiv preprint arXiv:2303.15647*.
- Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.
- Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang and Jie Tang. 2023. GPT understands, too. *AI Open*.
- Robert McCrae and Paul Costa. 1999. A Five-Factor theory of personality. *Handbook of personality: Theory and research* (pp. 139–153). Guilford Press, New York, NY.
- Angela Ramirez and Mamon Alsalihi and Kartik Aggarwal and Cecilia Li and Liren Wu and Marilyn Walker. 2023. Controlling Personality Style in Dialogue with Zero-Shot Prompt-Based Learning. *IWSDS (International Workshop on Spoken Dialogue Systems Technology) 2023*
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. *arXiv preprint arXiv:2212.10560*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean and William Fedus. 2022. Emergent Abilities of Large Language Models. *arXiv preprint arXiv:2206.07682*.
- Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, Zhiyuan Liu. 2021. Few-Shot Conversational Dense Retrieval. *arXiv:2105.04166*.
- Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li and Yu Qiao. 2023. LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention. *arXiv preprint arXiv:2303.16199*.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie and Ji-Rong Wen. 2023. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223*.
- Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou and Ji-Rong Wen. 2023. Large Language Models for Information Retrieval: A Survey. *arXiv preprint arXiv:2308.07107*.

## 1 A Appendix – Data Categories and Labels

economic\_indicators.10Y\_ust\_yield  
economic\_indicators.conf\_board\_leading\_indicators  
economic\_indicators.consumer\_confidence  
economic\_indicators.cpi  
economic\_indicators.cpi  
economic\_indicators.fed\_funds\_rate  
economic\_indicators.gdp  
economic\_indicators.id  
economic\_indicators.industrial\_production  
economic\_indicators.ism\_mfg  
economic\_indicators.ism\_non\_mfg  
economic\_indicators.personal\_consumption\_expenditure  
economic\_indicators.unemployment\_rate  
external\_data\_source  
market\_data.Communications\_services  
market\_data.Consumer\_Disc  
market\_data.Consumer\_Staples  
market\_data.Energy  
market\_data.Financials  
market\_data.Health\_Care  
market\_data.id  
market\_data.Industrials  
market\_data.IT  
market\_data.Materials  
market\_data.Real\_Estate  
market\_data.S&P500  
market\_data.Utilities  
news\_data\_source  
stock\_data.assets\_short\_term  
stock\_data.assets\_total  
stock\_data.bps\_book\_value\_per\_share  
stock\_data.cogs  
stock\_data.common\_shares\_outstanding  
stock\_data.company\_name  
stock\_data.date\_ex\_dividend  
stock\_data.dividend\_yield  
stock\_data.ebitda  
stock\_data.employees  
stock\_data.enterprise\_value  
stock\_data.eps\_forward\_12\_month  
stock\_data.eps\_trailing\_12\_month  
stock\_data.general\_industry\_classification  
stock\_data.id  
stock\_data.interest\_income  
stock\_data.interest\_income\_not\_banks  
stock\_data.inventories\_days\_held  
stock\_data.liabilities\_short\_term  
stock\_data.liabilities\_total  
stock\_data.long\_term\_debt  
stock\_data.market\_capitalization  
stock\_data.operating\_expense  
stock\_data.operating\_expense\_sg&a  
stock\_data.operating\_income  
stock\_data.price\_close  
stock\_data.price\_intraday\_high  
stock\_data.price\_intraday\_low  
stock\_data.ratio\_pe  
stock\_data.research\_and\_development  
stock\_data.sales\_or\_revenues  
stock\_data.shareholders\_equity  
stock\_data.ticker\_symbol  
stock\_data.total\_return



### 3 **B Appendix – QA Generation Examples**

#### 4 Base question generation from data categories and labels

5

##### **Few-Shot Prompt**

You are a everyday investor who is member of the general public not a professional investor. Generate a question that can be answered by and only by the following information source or database field. Use informal conversational tone. You MUST NOT include names of the available.

Information:

Chevron

stock\_data.research\_and\_development - RESEARCH AND DEVELOPMENT EXPENSE represents all direct and indirect costs related to the creation and development of new processes, techniques, applications and products with commercial possibilities.

Question: Does Chevron allocate funds for research and development?

Information:

news\_data - The news data is a data warehouse where we store all news scraped from Reuters. Each news item has keywords and date.

Question: Has there been any recent news about Nvidia?

Information:

stock\_data.ticker\_symbol - TICKER SYMBOL represents a symbol used to identify the company on the stock exchanges where it is listed.

Question: What is the ticker symbol for Apple Inc.?

Information:

news\_data - The news data is a data warehouse where we store all news scraped from Reuters. Each news item has keywords and date.

Question:

##### **Model Output**

Is there any news about the recent merger between Disney and Fox?

6

7

### Few-Shot Prompt

Here is some question: {nameVariable 20-25 decent familyFriendly}. Here is a rewrite of the question which is {UNCONSCIENTIOUSNESS} : {Oh gosh oh God I don't know. nameVariable is family friendly with a decent rating, also it has a price range of £20-25, I mean.}.

Here is some question: {nameVariable coffee shop Indian cheap average familyFriendly}. Here is a rewrite of the question which is {UNCONSCIENTIOUSNESS} : {Oh God err... I don't know. It's cheap and an Indian place, also nameVariable is a coffee shop, and it has an average rating. obviously, nameVariable isn't family friendly.}.

Here is some question: {nameVariable restaurant Japanese mediocre}. Here is a rewrite of the question which is {UNCONSCIENTIOUSNESS} : {Oh gosh I don't know buddy. nameVariable is a Japanese place with a damn mediocre rating, also it is a restaurant.}.

Here is some question: {nameVariable Italian cheap average riverside familyFriendly nearVariable}. Here is a rewrite of the question which is {UNCONSCIENTIOUSNESS} : {I don't know. nameVariable has an average rating, also nameVariable is damn cheap in riverside, also it is an Italian place kind of near nearVariable, and it is darn family friendly.}.

Here is some question: {nameVariable fast food low familyFriendly}. Here is a rewrite of the question which is {UNCONSCIENTIOUSNESS} : {I am not sure. nameVariable has like, a low rating, also it is a fast food place, and nameVariable isn't damn kid friendly friend.}.

Here is some question: {nameVariable coffee shop Japanese 20-25 high nearVariable}. Here is a rewrite of the question which is {UNCONSCIENTIOUSNESS} : {Oh God err... I don't know. I mean, nameVariable has a price range of £20-25 with a high rating, and it is a coffee shop, also nameVariable is a Japanese place near nearVariable.}.

Here is some question: {nameVariable Chinese high average nearVariable}. Here is a rewrite of the question which is {UNCONSCIENTIOUSNESS} : {Yeah, oh God I don't know. Err... nameVariable has an average rating, and it is a Chinese place and expensive near nearVariable.}.

Here is some question: {nameVariable cheap excellent familyFriendly nearVariable}. Here is a rewrite of the question which is {UNCONSCIENTIOUSNESS} : {Yeah, I am not sure. nameVariable is cheap and kid friendly near nearVariable with a damn excellent rating!}.

Here is some question: {nameVariable pub Indian 20-25 nearVariable}. Here is a rewrite of the question which is {UNCONSCIENTIOUSNESS} : {Oh God oh gosh I don't know. nameVariable is an Indian place, nameVariable is a pub, and it is kind of near nearVariable, also it has a price range of £20-25.}.

Here is some question: {nameVariable French moderate}. Here is a rewrite of the question which is {UNCONSCIENTIOUSNESS} : {Oh gosh I mean, I am not sure. nameVariable is damn moderately priced and a French place.}.

Here is some question: {What does Apple do?}. Here is a rewrite of the question which is {UNCONSCIENTIOUSNESS} : {

### Model Output

Oh gosh, erm... you know... what is it that Apple does again?